THE UTILITY OF *atp*B GENE SEQUENCES IN RESOLVING PHYLOGENETIC RELATIONSHIPS: COMPARISON WITH *rbc*L AND 18S RIBOSOMAL DNA SEQUENCES IN THE LARDIZABALACEAE[1]

Sara B. Hoot,[2] Alastair Culham,[3] and Peter R. Crane[4]

ABSTRACT

The chloroplast gene *atp*B was sequenced for seven genera of the Lardizabalaceae and three outgroup taxa to assess its utility as a source of phylogenetic information. The resulting phylogenetic tree was compared with trees based on 18S nuclear ribosomal DNA and *rbc*L (chloroplast DNA) sequences, as well as a combination of all data (*atp*B, 18S, and *rbc*L) for the same taxa. Sequence divergence values, statistics related to patterns of character transformation, and indices measuring homoplasy and branch support were also compared. The topology of the trees derived from *atp*B, 18S, and a combination of all three sequence data sets were largely congruent. All phylogenies, with the exception of the tree derived from *rbc*L data, supported the monophyly of the Lardizabalaceae. All indicators of nucleotide substitution rate suggest that *rbc*L is the least conserved, *atp*B is intermediate, and 18S is the most conserved of the three genes sequenced. Measures of homoplasy also indicate that the *rbc*L tree is less strongly supported than those based on *atp*B, 18S, or a combination of *atp*B, 18S, and *rbc*L sequence data.

Phylogenetic analyses of higher-level plant groups using DNA sequence data have been based most often on the chloroplast gene *rbc*L or, less frequently, on 18S nuclear ribosomal DNA (18S nrDNA). Few phylogenetic studies have used other gene sequences across a broad range of taxa, and still fewer have compared results from two or more gene sequences for the same taxa (e.g., Baldwin, 1992; Johnson & Soltis, 1994; Olmstead & Sweere, 1994). As part of an intensive systematic study of phylogenetic relationships among basal eudicots (Ranunculidae and "lower" Hamamelididae), we have further developed the chloroplast gene, *atp*B, as a new source of phylogenetically informative data (Ritland & Clegg, 1987). Here, we apply the *atp*B gene to resolve phylogenetic relationships in the angiosperm family Lardizabalaceae and compare the results with those based on *rbc*L and 18S nrDNA for the same taxa. Cladograms based on the three genes are evaluated and compared in terms of their resolution and congruence, as well as various measures of phylogenetic signal, transition/transversion bias, sequence divergence, and homoplasy. Based on these data, we discuss the effectiveness of using each gene for phylogenetic studies at the generic level and above.

The Lardizabalaceae (Ranunculidae, sensu Takhtajan, 1987—"ranunculids") are a family of twining (rarely erect) shrubs found in temperate areas of Eastern Asia and South America. The family is characterized by alternate, palmate (rarely pinnate) compound leaves; regular unisexual flowers; six overlapping or valvate sepals (three in *Akebia*); staminodia or petals small or absent; three

[1] We thank R. Olmstead and S. Lanyon for helpful comments on this manuscript, D. Nickrent and E. A. Zimmer for ribosomal DNA amplification and sequencing primers, and G. Zurawski for *rbc*L sequencing primers. We are grateful to the following for providing us with leaf material: R. McBeath, Royal Botanic Garden, Edinburgh; A. Reznicek, University of Michigan; T. Lammers, Field Museum of Natural History; M. Maunder and T. Upson, Royal Botanic Gardens, Kew; H. Qin, Chinese Academy of Sciences, Beijing; and E. van Jaarsveld, Kirstenbosch Botanical Garden. We also thank J. Hall for technical advice and A. Thurmond for laboratory assistance. This work was supported in part by NSF grants DEB-9020237 to PRC and DEB-9306533 to SBH.

[2] Department of Biological Sciences, University of Wisconsin–Milwaukee, Lapham Hall, P.O. Box 413, Milwaukee, Wisconsin 53201, U.S.A.

[3] Department of Botany, School of Plant Sciences, University of Reading, Reading, RG6 2AS, U.K.

[4] Departments of Botany and Geology, Field Museum, Chicago, Illinois 60605, U.S.A.

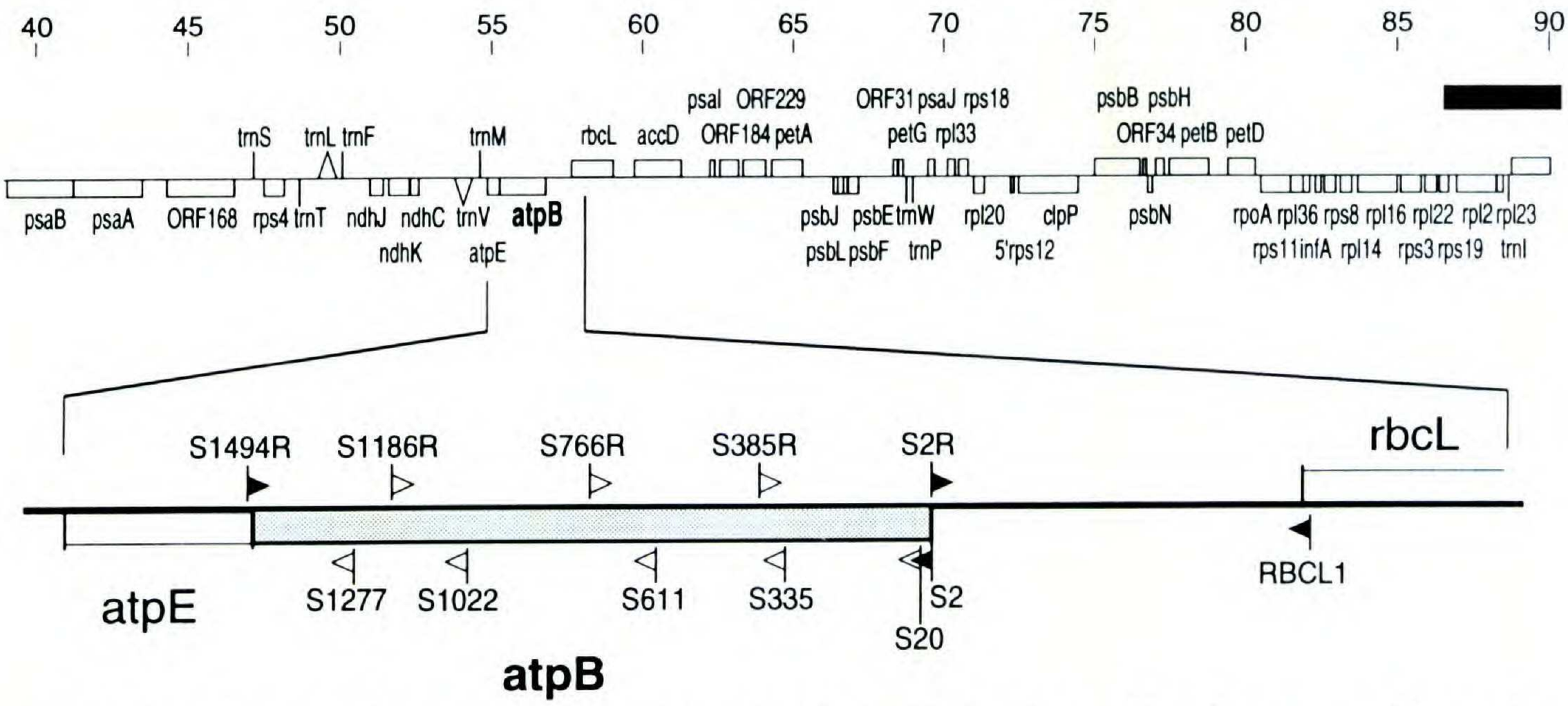


FIGURE 1. Location of the *atp*B gene in relation to a portion of the tobacco chloroplast genome. Top numbers indicate tobacco coordinate units × 1000 (TCU; Shinozaki et al., 1986). Line below the TCU scale maps a portion of the single copy region and inverted repeat of the tobacco chloroplast genome from TCU 40,000 to TCU 90,000. Black bar at approximately 87,000 TCUs is the beginning of the inverted repeat. Genes above the tobacco genome map (Shinozaki et al., 1986) are transcribed from left to right; those below the line are transcribed from right to left. The bold line beneath the tobacco genome map illustrates the bracketed region above in more detail. Filled in arrows indicate the location and direction of the amplification primers for the *atp*B gene. Hollow arrows indicate the location and direction of the internal sequencing primers.

carpels (up to nine in *Akebia*); numerous ovules (four, by abortion, in *Boquila*) with laminar placentation (submarginal in *Decaisnea* and *Sinofranchetia*).

Previous phylogenetic analyses based on *rbc*L sequence data suggest that the Lardizabalaceae occupy a key and potentially basal position in the evolution of the Ranunculidae, but the family was represented by only the genus *Akebia* (Chase et al., 1993). This molecular study was undertaken to clarify intergeneric relationships in the family, to test the results of previous morphologically based cladistic analyses (Loconte & Estes, 1989), and to provide a more secure basis for representing the family in our ongoing investigation of basal eudicot radiation. Of particular interest are the phylogenetic positions of *Decaisnea* and *Sinofranchetia* (the only Lardizabalaceae genera with sub-marginal placentation) with respect to the other genera and the relationships of the two dioecious South American genera, *Boquila* and *Lardizabala*, the only representatives of the family outside eastern Asia.

THE *atp*B GENE

The *atp*B gene is located in the large single-copy region of the chloroplast genome contiguous with the *atp*E gene and downstream from the *rbc*L gene, from which it is separated by an approximately 900 bp intergenic spacer region (Fig. 1). The *atp*B gene encodes the β subunit of ATP synthase (other subunits are encoded in either the chloroplast or the nuclear genomes). ATP synthase has a highly conserved structure that couples proton translocation across membranes with the synthesis of ATP (Zurawski et al., 1982; Gatenby et al., 1989). Previous to this study, the chloroplast *atp*B gene had been sequenced for approximately ten genera, representing a diverse range of plants (e.g., *Chlamydomonas*, *Marchantia*, *Spinacia*, *Nicotiana*, and *Oryza*).

Many features of the *atp*B gene suggest that it may be valuable for comparative sequence studies at higher taxonomic levels. It is short enough (1497 bp) for ease of sequencing but long enough to be potentially phylogenetically informative, given broadly comparable rates of evolution to *rbc*L. The evolutionary rate is conserved and K_s (a measure of the rate of synonymous nucleotide substitution in the gene; Li et al., 1985; Wolfe, 1991) between rice and tobacco is 0.62, indicating a rate of evolution very similar to that found for *rbc*L (K_s = 0.63; Wolfe, 1991). There are no reported insertions and deletions in the *atp*B gene, the gene does not contain introns, and *atp*B sequences are readily aligned.

MATERIALS AND METHODS

DNA EXTRACTION AND AMPLIFICATION

The seven genera of Lardizabalaceae and three outgroup genera used in this study are indicated

TABLE 1. Species sequenced, bases sequenced for each gene (only missing data of ≥10 nucleotides are noted as gaps), total number of variable sites missing from each sequence (listed in parentheses), sources of sequences (SH—Sara Hoot, AC—Alastair Culham), and voucher information with notations on nature of material (fresh, silica-dried, or herbarium). Accession numbers (GSDB) are in the following order: *atp*B, 18S, and *rbc*L.

Species	*atp*B Bases sequenced and source	18S Bases sequenced and source	*rbc*L Bases sequenced and source	Accession information and location of voucher
Decaisnea fargesii Franchet	25–358, 390–1493 (5) SH	39–1710 (1) SH	31–1261, 1276–1428 (3) SH	L37926, L37907, L37916 Reznicek 9236, MICH (fresh)
Lardizabala biternata Ruíz & Pavón	57–1493 (0) SH/AC	39–1710 (0) SH	31–1428 (0) SH/AC	L37929, L37910, L37919 Edinburgh 271032, F (fresh)
Holboellia latifolia Wallich		39–1710 (1) SH		L37909 Culham 193, RNG (silica-dried)
Holboellia latifolia Wallich	57–1493 (2) SH/AC		31–1428 (1) SH/AC	L37928, L37918 Edinburgh 715977, F (fresh)
Sinofranchetia chinensis Hemsley	61–1493 (1) SH/AC	39–1710 (3) SH	31–1300 (15) SH	L37931, L37912, L37921 Edinburgh 831635, F (fresh)
Stauntonia hexaphylla (Thunb.) Decne.	25–1493 (0) SH	39–1411, 1433–1710 (2) SH	31–1428 (1) SH	L37932, L37913, L37922 Lammers 8465, F (silica-dried)

TABLE 1. Continued.

Species	*atp*B Bases sequenced and source	18S Bases sequenced and source	*rbc*L Bases sequenced and source	Accession information and location of voucher
Akebia quinata Decne.			29–1428 (1) Qiu	L12627 Qiu 91020, NCU
Akebia quinata Decne.	68–364, 388–1493 (5) SH	39–1710 (0) SH	31–300 SH	L37924, L37905 Hoot 925, F (fresh)
Boquila trifoliata Decne.	19–368, 381–1493 (2) SH/AC	39–595, 615–1710 (1) SH	31–1428 (5) SH/AC	L37925, L37906, L37915 Gardner & Knees 4088, RNG (1988) (Herbarium specimen)
Dicentra eximia Torrey	25–379, 401–1493 (4) SH	39–1170, 1191–1710 (2) SH	34–1428 (3) SH	L37927, L37908, L37917 Reznicek 9756, MICH (fresh)
Nandina domestica Thunb.	60–339, 388–1493 (7) SH	39–1710 (1) SH	67–780, 813–1428 (8) SH	L37930, L37911, L37920 Matthaei Bot. Gard., Hoot 922 F (fresh)
Tinospora caffra Miers	59–1493 (1) SH	39–1710 (0) SH	31–1315, 1329–1421 (1) SH	L37933, L37914, L37923 Jaarsveld 2131, NBG (silica-dried)

TABLE 2. Location (position based on *atp*B sequence for *Spinacia*, Zurawski et al., 1982) and base composition of amplification and internal sequencing primers for the chloroplast gene, *atp*B. All primers were designed by S. Hoot.

	Amplification primers
RBCL1	5′ GAA TCC AAC ACT TGC TTT AGT CTC T (for amplification of the spacer region between *atp*B and *rbc*L as well as the *atp*B gene)
S2	5′ TAT GAG AAT CAA TCC TAC TAC TTC T (amplifies the *atp*B gene only)
S1494R	3′ TCA GTA CAC AAA GAT TTA AGG TCA T
	Internal sequencing primers, forward direction
S20	5′ CTT CTG ATC CTG GGG TTT CCA CAC T
S335	5′ ACG TGC TTG GGG AGC CTG TTG ATA A
S611	5′ AAC GTA CTC GTG AAG GAA ATG ATC T
S1022	5′ CGA CAT TTG CAC ATT TAG ATG CTA C
S1277	5′ AAA TTG AGC GTT TCT TAT CAC AAC C
	Internal sequencing primers, reverse direction
S2R	5′ AGA AGT AGT AGG ATT GAT TCT CAT A
S385R	5′ GCG CAG ATC TAT GAA TAG GAG ACG T
S766R	5′ TAA CAT CTC GGA AAT ATT CCG CCA T
S1186R	5′ TGT CCT GAA GTT CTT TGT AAC GTT G
S1494R	same as amplification primer above

in Table 1, along with sequencing, accession, and voucher information. Total cellular DNA was isolated from fresh, silica-dried, or herbarium material according to the miniprep method of Doyle & Doyle (1987). In some cases (*Boquila trifoliata*, *Sinofranchetia chinensis*, *Stauntonia hexaphylla*), DNA was further purified and concentrated after extraction using GeneClean (GeneClean, Bio 101, Inc.).

A segment of double-stranded DNA containing most of the coding sequence (approximately 1474 bp) for the *atp*B gene was amplified using the polymerase chain reaction (PCR). Three amplification primers were designed using *atp*B sequences available from GenBank for spinach, tobacco, pea, sweet potato, maize, and wheat. Two alternate 25-mer 5′ primers were used. One is located at the 5′ end of the *rbc*L coding sequence (nucleotide positions 15–39 in tobacco). It has the opposite orientation from *rbc*L but the same orientation as *atp*B, and includes the intergenic spacer region between *rbc*L and *atp*B in the amplification product (Table 2, Fig. 1). The other primer begins with the first base upstream of the *atp*B start codon and includes the first 24 bases of the *atp*B coding sequence (positions 1–24 in spinach). The 3′ 25-mer amplification primer, S1494R, is located at the 5′ end of the adjacent gene, *atp*E (positions 1–25 of the *atp*E gene in spinach), and has the opposite orientation as the *atp*B gene (Table 2).

Two alternative protocols differing only in $MgCl_2$ concentration and annealing temperature were used to amplify *atp*B. In the first protocol, the reaction mixture contained (either final concentrations or amounts in a 100 μl reaction): 10 mM Tris-HCl, pH8.3, 50 mM KCl, 1.5 mM $MgCl_2$, 0.2 mM of each deoxyribonucleotide triphosphate (dNTP), 0.5 μM of each amplification primer, 2.5U Taq Polymerase, 0.3–2.0 μL template DNA (depending on concentration). To prevent evaporation during thermal cycling, a drop of mineral oil was added to each reaction mixture. The sample was then placed in a thermocycler (M. J. Research, Inc., Cambridge, Massachusetts) with the following cycling parameters: premelt at 92°C for 3 min.; 30 cycles, each consisting of a denaturation step at 92°C for 1 min., annealing step at 55°C for 1 min., and an extension step at 72°C for 3 min.; followed by a final extension step of 72°C for 7 min. The alternative protocol included the following modifications: the concentration of $MgCl_2$ was doubled and the annealing step temperature lowered to 50°C. In most cases one of these two protocols produced amplification product. In those cases where yield was still very weak, further amplification directly from the Gene-Cleaned amplification product often produced increased yields.

In most cases, the amplification primers for the chloroplast gene *rbc*L were those described in Olmstead et al. (1992). In *Sinofranchetia*, no amplification product could be obtained using these primers. Substituting a 3′ primer located at position

1300 in the tobacco *rbc*L sequence and using an 18S amplification protocol (Nickrent, 1993) resulted in high yields (Table 1). 18S nrDNA was amplified by using either the primers and protocols of Nickrent (Nickrent, 1993; Nickrent & Starr, 1993) or those of Hamby et al. (1988).

Particular attention was given to purification of PCR products to avoid superimposed sequences that can result from priming by the amplification primers as well as the internal sequencing primer during double-standed sequencing. Samples were run on a 2% low melt agarose gel (NuSieve GTG) with 1% TAE buffer and ethidium bromide, bands were visualized by means of UV illumination, then removed as gel plugs. To remove agarose and concentrate the PCR product, gel plugs were melted at 65°C for approximately 10 min., then further purified and concentrated with glass milk (GeneClean). This procedure proved especially important to obtain high-quality sequences for *rbc*L and 18S nrDNA but amplification products for *atp*B frequently gave excellent results when sequenced after using only the GeneClean purification process.

DOUBLE-STRANDED SEQUENCING

The purified dsPCR product was sequenced directly with the dideoxy-termination method and Sequenase T7 DNA polymerase (US Biochemical) using the protocol of Thien (1990) with the following modifications: the addition of 1% acetamide to the annealing reaction and an incubation temperature of 46–47°C for the termination step. Internal sequencing primers for *atp*B are shown in Table 2 and Figure 1. Sequencing of *rbc*L used a combination of internal primers, kindly provided by G. Zurawski (DNAX Research Institute, Palo Alto, California) and a few primers designed specifically for this study (positions and sequences available from SBH). The 18S internal primers used to date were generously furnished by D. Nickrent (Nickrent & Starr, 1993) or E. A. Zimmer (Hamby et al., 1988).

Aliquots of the sequencing reactions were loaded on two 60 × 33 cm field-gradient 6% polyacrylamide gels and subjected to electrophoresis overnight (short run—600V, long run—1100V). After transferring to 3MM Whatman paper, the gels were vacuum-dried for approximately one hour at 80°C and exposed to autoradiography film for 1–3 days. Typical autoradiographs from gels run in this manner yielded 300 to 350 readable bases. All sequences used in this study are available from GSDB or directly from the senior author. Consistently, *atp*B proved the easiest to sequence of the three genes examined: it amplified readily, suffered the least from multiple banding patterns when the PCR product was not gel-purified, and often yielded readable sequences 300–350 bp from the internal primers.

QUALITY CONTROL OF SEQUENCE DATA

Sequence comparisons for the genes *atp*B, 18S, and *rbc*L included 1468, 1671, and 1397 bp, respectively (Table 1). Both strands of DNA were sequenced for both *atp*B and *rbc*L with approximately 80% overlap. Both strands were also sequenced for 18S, but with much less overlap between the two directions (30–40%). The sequences were read from the autoradiographs, recorded on a data sheet, entered into MacClade (Maddison & Maddison, 1992), then printed and rechecked from the autoradiographs for errors.

Alignment problems (caused by compressions) for *atp*B and *rbc*L often could be rectified by reading the opposite strand. Within the *atp*B gene, the following regions were susceptible to compressions: positions 47–53, 875–879, and 1455–1457. There were several regions in the 18S nrDNA sequences where alignment was impossible because of compressions or base insertion/deletion events. These regions were deleted from the data matrix and are located at the following positions in relation to the soybean 18S sequence (Eckenrode et al., 1985): 224–231, 667–670, 710, 738, 1174–1175, 1366, and the very end of the amplified region, 1711–1761. Sequence divergence values (described below) calculated between pairs of sequences excluded these problematic regions or positions. The possibility of PCR-generated anomalous sequences was checked by comparison of sequences from closely related taxa. Sequences furnished by other labs were also checked for inconsistencies both by comparison with other closely related taxa and occasional duplicate sampling of the same genus (e.g., the *rbc*L and 18S sequences for *Akebia*, and the *rbc*L sequence for *Dicentra*).

DATA ANALYSIS

Phylogenetic analyses were performed using PAUP 3.1 (Swofford, 1993) using the branch-and-bound search option (with collapse of zero-length branches) to assure recovery of the most parsimonious trees. PAUP was also used to perform bootstrap analysis with 1000 replications using the branch-and-bound search option (Felsenstein, 1985). The decay indices (the number of steps that must be added to the minimal-length tree before a

TABLE 3. Comparison of data sets from *atp*B, 18S nrDNA, *rbc*L, and a combination of all data. Numbers in parentheses indicate number of informative three-state characters (excluding those where two of the three states were autapomorphies). The % Ts = unambiguous transitions/unambiguous changes × 100 and was calculated for each gene from one of the most parsimonious trees. g_1 is a measure of the skewness of the distribution of 100,000 randomly generated trees. Tree length (TL) was calculated including uninformative characters; consistency index (CI) and rescaled consistency index (RC) were calculated excluding uninformative characters. RI = retention index.

Gene	Variable sites	Informative sites	Binary characters	3-state characters	4-state characters	%Ts	g_1	Number of trees	TL	CI	RI	RC
*atp*B	143	42	36	11 (6)	0	69	−1.36	3	172	0.73	0.74	0.54
18S	76	25	21	6 (4)	0	66	−1.19	3	88	0.81	0.85	0.69
*rbc*L	172	56	47	21 (9)	0	56	−0.59	1	225	0.64	0.68	0.44
Combined data	391	123	104	38 (19)	0	61	−0.82	2	493	0.67	0.70	0.47

clade collapses) were computed for all trees using the heuristic search option (Donoghue et al., 1992). A tree length distribution of 100,000 randomly sampled trees was generated for each of the *atp*B, 18S, and *rbc*L data matrices using the "random trees" selection of PAUP. This distribution was analyzed for skewness as an estimate of nonrandom structure in the sequences (Hillis & Huelsenbeck, 1992). Sequence divergence values were computed as the proportion of divergent sites from direct pairwise comparisons of the sequence data.

There has been much discussion about the advantages and disadvantages of analyzing independent data sets separately, combining independent data sets prior to phylogenetic analyses, or analyzing them separately and then applying consensus methods (see Bull et al., 1993, for a review). There is clearly a strong argument for combining data sets, especially in cases where there is substantial homoplasy and the phylogenetic signal in a particular data set is insufficient to resolve certain branching patterns in a tree (Kluge, 1989; Barrett et al., 1991; Olmstead & Sweere, 1994). For this reason, we have chosen to analyze a combination of all the data sets (*atp*B, 18S, and *rbc*L) as well as each data set separately.

Alternative tree topologies and resultant changes in tree length were explored using MacClade 3.0 (Maddison & Maddison, 1992). MacClade was also used to calculate character transformations of various types for each sequence tree (for example, transition/transversion bias and the number of changes at different codon positions).

Outgroup taxa for the Lardizabalaceae in all analyses were selected based on the results of several previous phylogenetic analyses of the Ranunculidae (sensu Takhtajan, 1987). A cladistic analysis based on traditional data placed Berberidaceae, Menispermaceae, and the Ranunculineae (Ranunculaceae plus Papaveraceae) as potential sister taxa to the Lardizabalaceae (Loconte & Stevenson, 1991). However, preliminary analyses of *atp*B and *rbc*L data (analyzed as separate and combined data sets) with extensive sampling of the ranunculids, consistently place the Ranunculaceae as the most derived family of the Ranunculidae. Representatives from two families, Menispermaceae and Berberidaceae, are the basal members of a large clade that is resolved as the sister group to the Lardizabalaceae (Hoot & Crane, work in progress). The Papaverales are resolved as relatively basal to the Lardizabalaceae and other ranunculids (Chase et al., 1993; Hoot & Crane, work in progress). Therefore, in this paper *Dicentra eximia* (Fumariaceae, Papaverales) was used to root the phylogenetic analyses with *Tinospora* (Menispermaceae) and *Nandina* (Berberidaceae) included as additional outgroup taxa (Table 1).

RESULTS

PHYLOGENETIC ANALYSES

Table 3 lists the number of variable positions, informative characters (after removal of autapomorphies), and binary, three- and four-state characters for each gene.

Analysis based on the *atp*B data resulted in three equally parsimonious trees; *Stauntonia*, *Akebia*, and *Holboellia* are unresolved due to lack of variable sites. These trees were based on 143 variable sites (Table 3; 42 informative characters) with a tree length (TL) = 172, a consistency index excluding autapomorphies (CI) = 0.73 (Kluge & Farris, 1969), and a retention index (RI) = 0.74

(Farris, 1989). The strict consensus cladogram derived from the three trees is presented in Figure 2. The monophyly of the Lardizabalaceae is well supported, with 19 base substitutions, a bootstrap value of 100%, and a decay index of 11. *Sinofranchetia* and *Decaisnea* are basal within the family. The remaining five genera form two clades, each of which is also supported by morphological characters (Hoot, Culham & Crane, work in progress). The clade consisting of the two South American genera, *Boquila* and *Lardizabala*, is only weakly supported (one base substitution, a bootstrap value of 64%, and a decay index of one) in contrast to stronger support for the clade comprising the Asian genera, *Stauntonia*, *Akebia*, and *Holboellia* (four base substitutions, a bootstrap value of 94%, and a decay index of three).

The 18S nrDNA data matrix, consisting of 76 variable sites (25 informative sites), resulted in three most parsimonious trees with a TL = 88, CI (excluding autapomorphies) = 0.81, and RI = 0.85. One of the shortest trees is presented in Figure 2 (the branch that collapses in the strict consensus tree is indicated with dotted lines). As in the *atp*B tree, the monophyly of the Lardizabalaceae with respect to the three outgroups is well supported with nine nucleotide changes, a bootstrap value of 99%, and a decay index of seven. The 18S tree is congruent with the *atp*B tree in other respects as well: *Sinofranchetia* and *Decaisnea* are basal within the family and the clade consisting of *Lardizabala*, *Boquila*, *Akebia*, *Stauntonia*, and *Holboellia* is recognized but with less internal resolution.

The *rbc*L data consisting of 172 variable sites (56 informative sites) resulted in one fully resolved most parsimonious tree (Fig. 2). The TL = 225, CI (excluding autapomorphies) = 0.64, and RI = 0.68. However, unlike the results from *atp*B and 18S, the *rbc*L sequence data do not support the monophyly of the Lardizabalaceae, placing *Sinofranchetia* with the outgroup *Dicentra* in a basal position with respect to all other genera. The *rbc*L tree is congruent with the *atp*B tree in placing *Decaisnea* as the sister genus to the clade consisting of *Lardizabala*, *Boquila*, *Akebia*, *Stauntonia*, and *Holboellia*, although the pattern of relationships among these five genera is different (the clade containing *Lardizabala* and *Boquila* is not recognized).

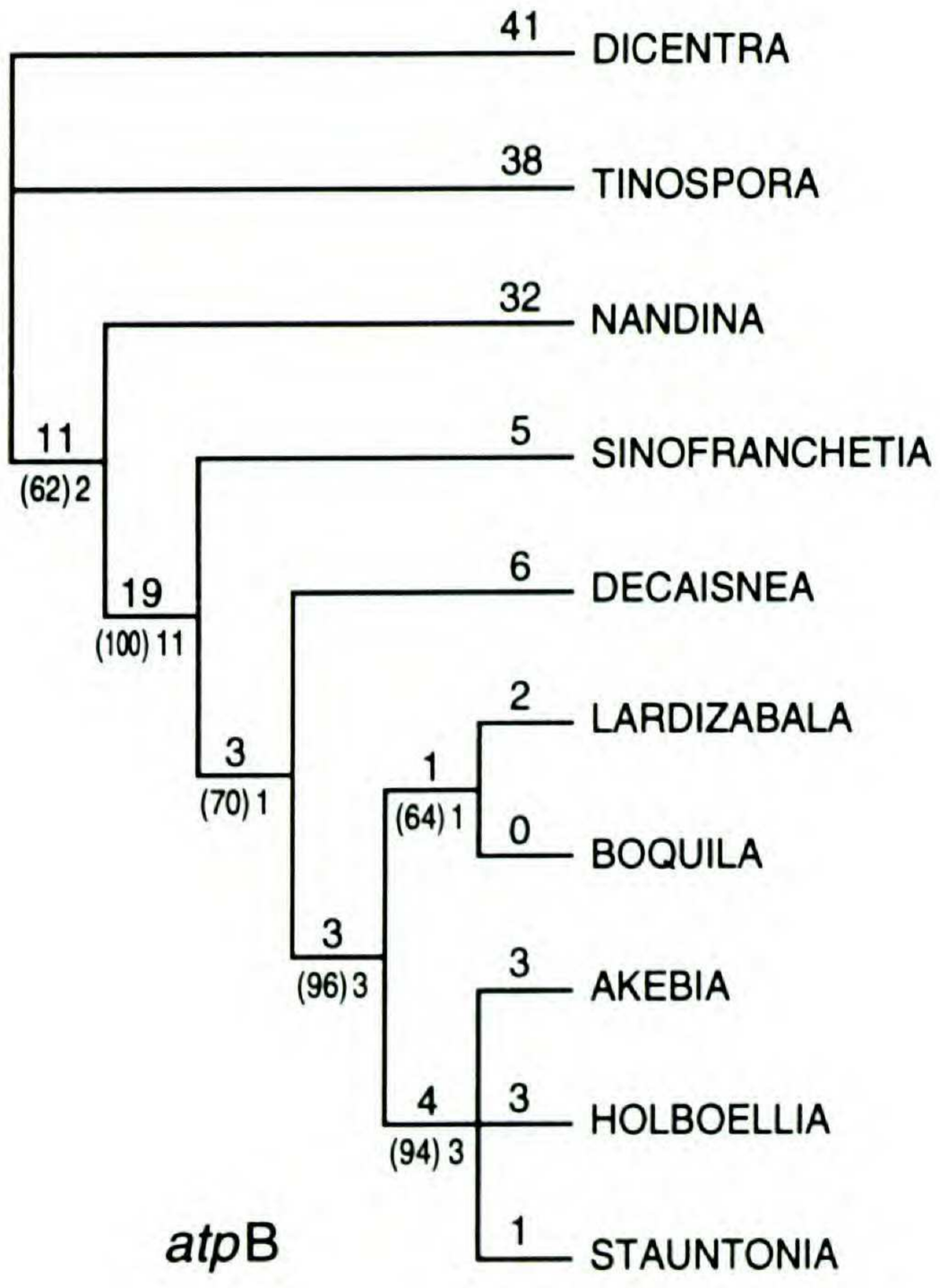
A further analysis performed using a combination of all three data sets (*atp*B, *rbc*L, and 18S nrDNA) resulted in two trees derived from 391 variable sites (123 informative sites) with a TL = 493, CI (excluding autapomorphies) = 0.67, and RI = 0.70. The clade that collapses in a strict consensus tree of the two most parsimonious trees is shown with dotted lines (Fig. 2). Excluding relationships among outgroups, the consensus tree based on all data is most similar to the tree based on *atp*B data alone. The monophyly of the Lardizabalaceae is again strongly supported with 37 nucleotide changes, a bootstrap value of 100%, and a decay index of 14. The combined data also support the clades consisting of (*Lardizabala*, *Boquila*, (*Akebia*, *Stauntonia*, and *Holboellia*,)) but with higher bootstrap values and decay indices (Fig. 2).

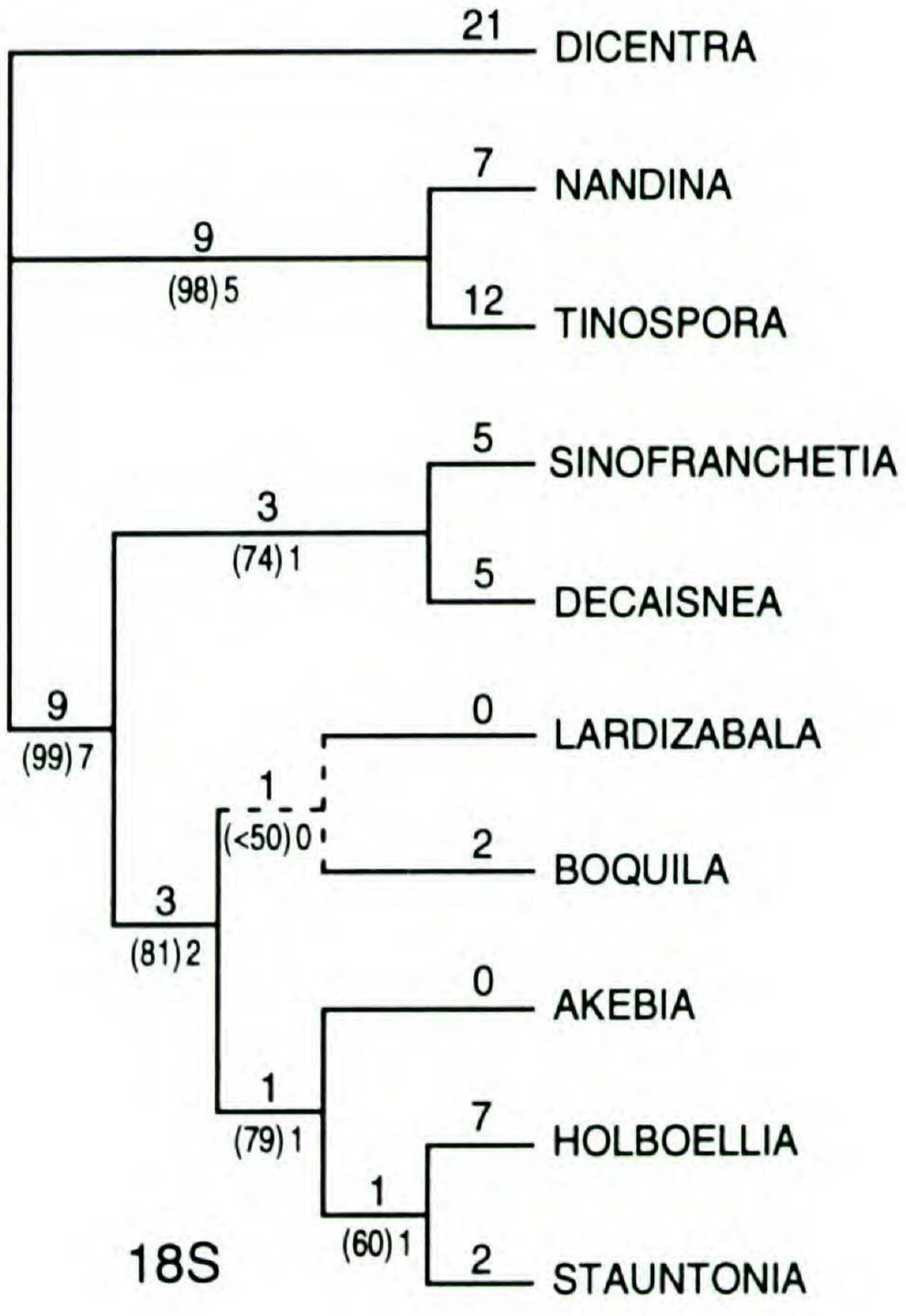
CHARACTER TRANSFORMATIONS

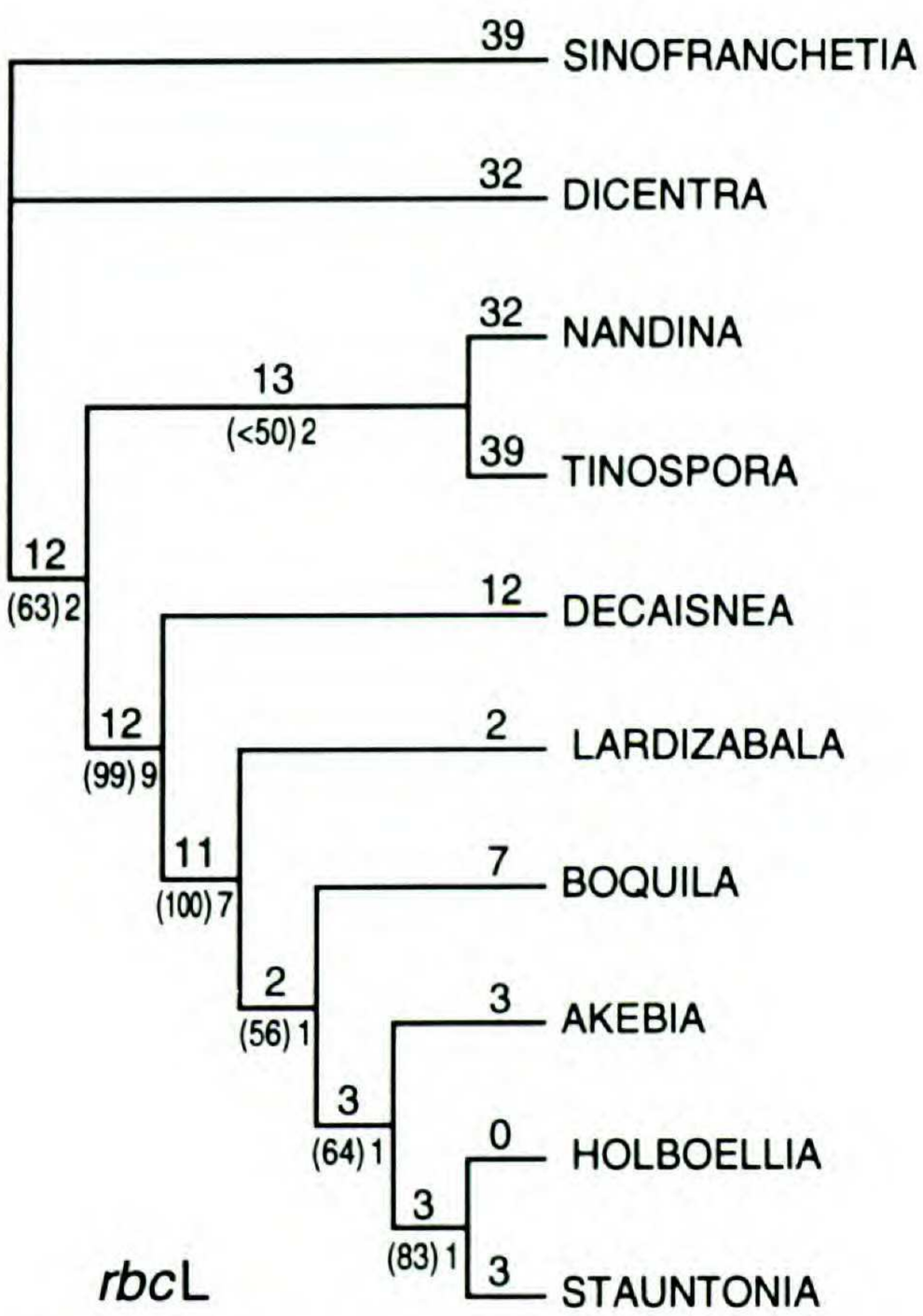
The number of substitutions inferred for each nucleotide position was calculated over the trees derived from the three sequence data sets using MacClade 3.0 (Maddison & Maddison, 1992). For both *atp*B and *rbc*L, substitutions occur fairly uniformly across the gene (Fig. 3). Even considering the exclusion of some nucleotide positions due to compressions or insertion/deletion events (see above), the nucleotide substitutions are much less evenly distributed across the 18S gene. This is probably due to constraints imposed by the secondary structure of 18S nrDNA (Gutell & Woese, 1990; Nickrent & Sargent, 1991; Dixon & Hillis, 1993). There is variation in the number of steps/site found for each gene, with a high of four steps in *atp*B to seven in *rbc*L.

Most of the unambiguous changes (142 of 172 changes or 83%) in the most parsimonious *atp*B trees occur at third-position sites contrasted with 15 unambiguous changes each at first and second positions. A similar pattern is found with the *rbc*L tree: 41 changes in the first-position, 21 in second-position, and 167 in third-position sites (73%). The CIs (including autapomorphies) for the various positions are high for both genes, even at third-position sites: 1.0 (*atp*B) and 0.73 (*rbc*L) for first positions, 0.93 and 0.90 for second positions, and 0.88 and 0.86 for third positions.

There are approximately twice as many transitions (94–95) as transversions (42–43) when calculated over the three most parsimonious *atp*B trees. A similar proportion of transitions to transversions occurs with the 18S data, 51–52 transitions to 25–26 transversions. However, the proportion is more even when calculated over the most parsimonious *rbc*L tree, 64 transitions to 51 transversions.







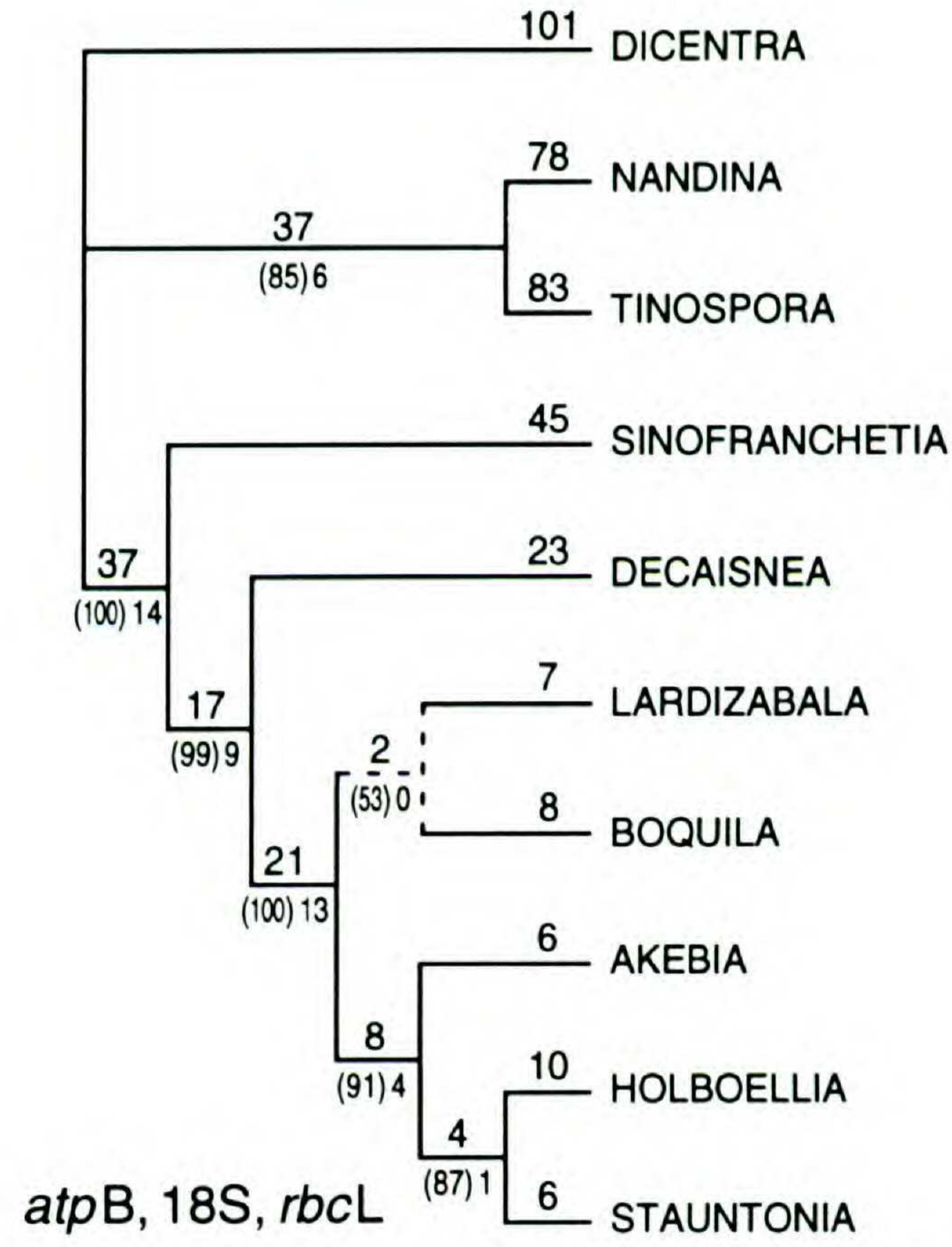
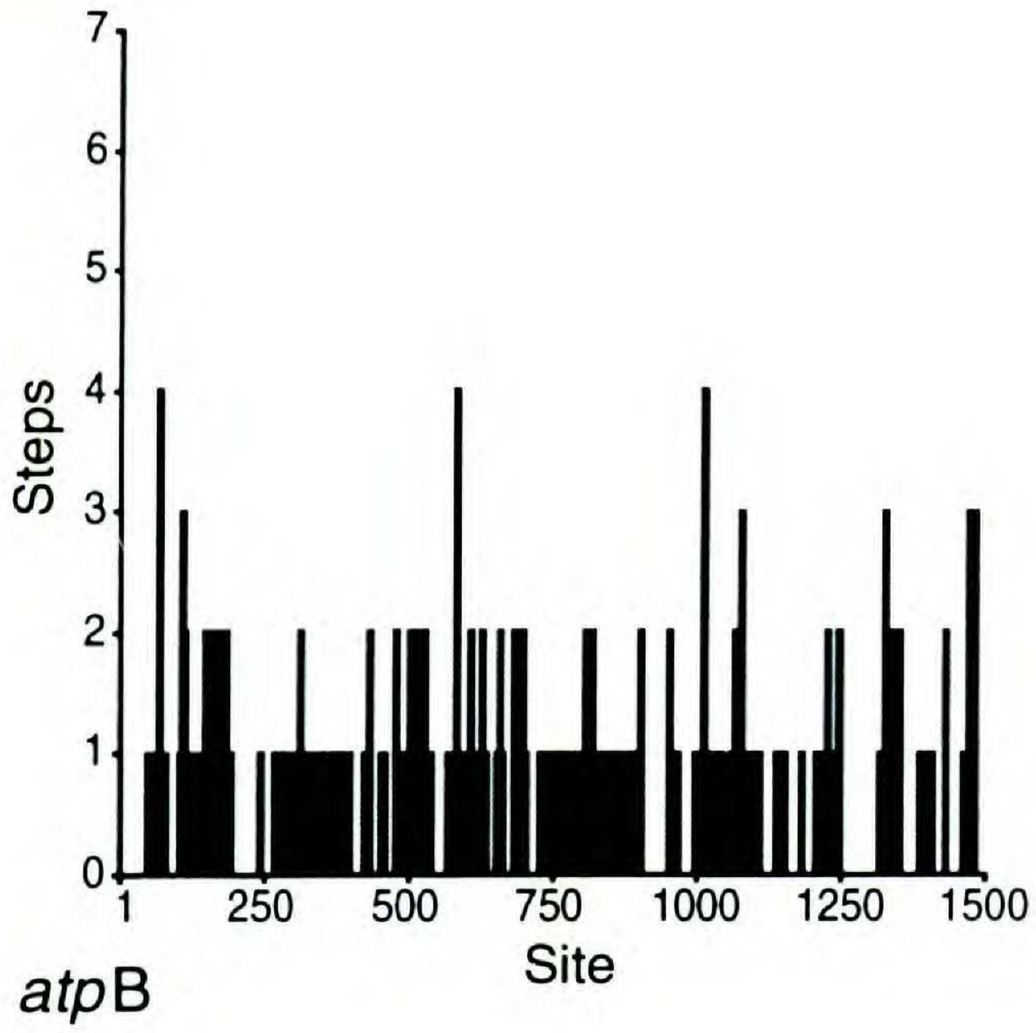
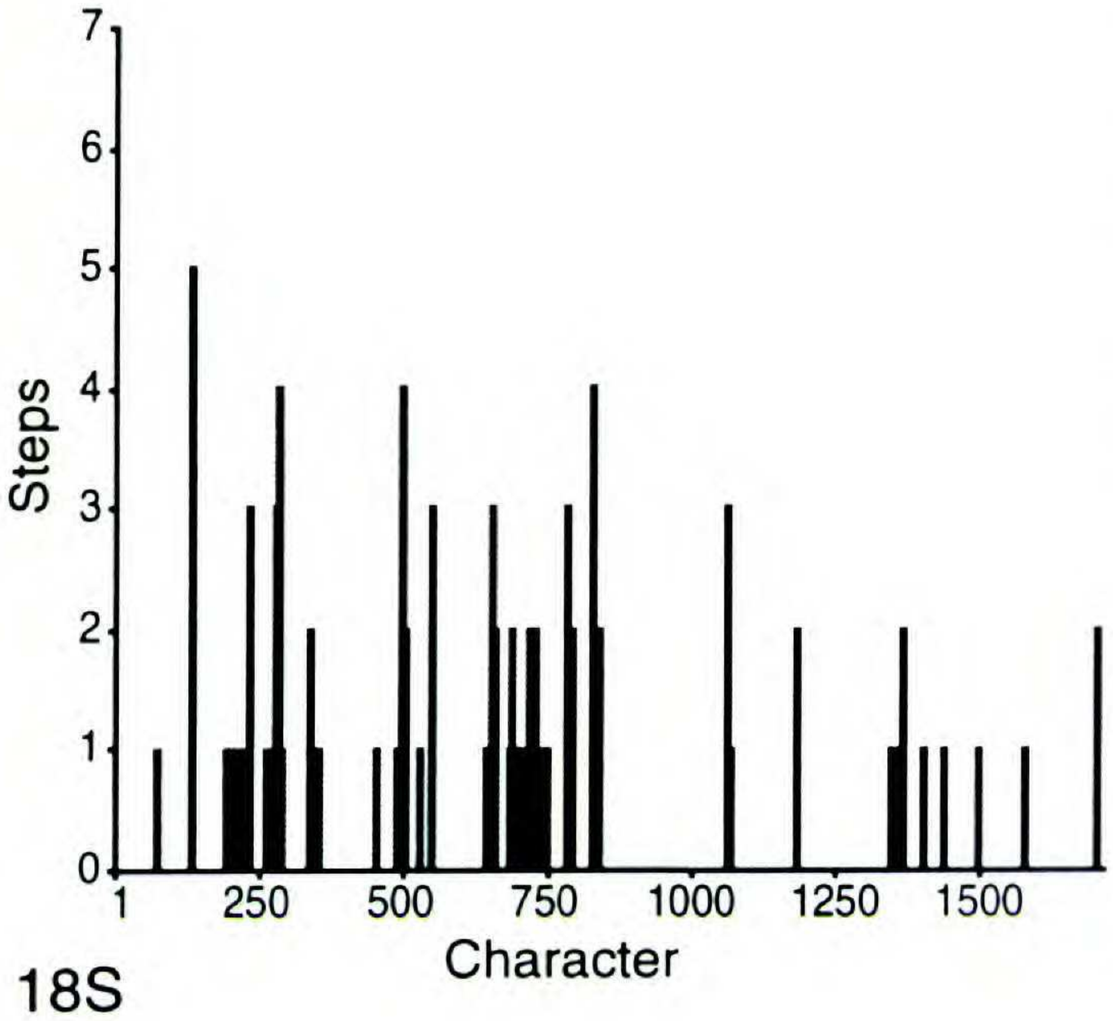
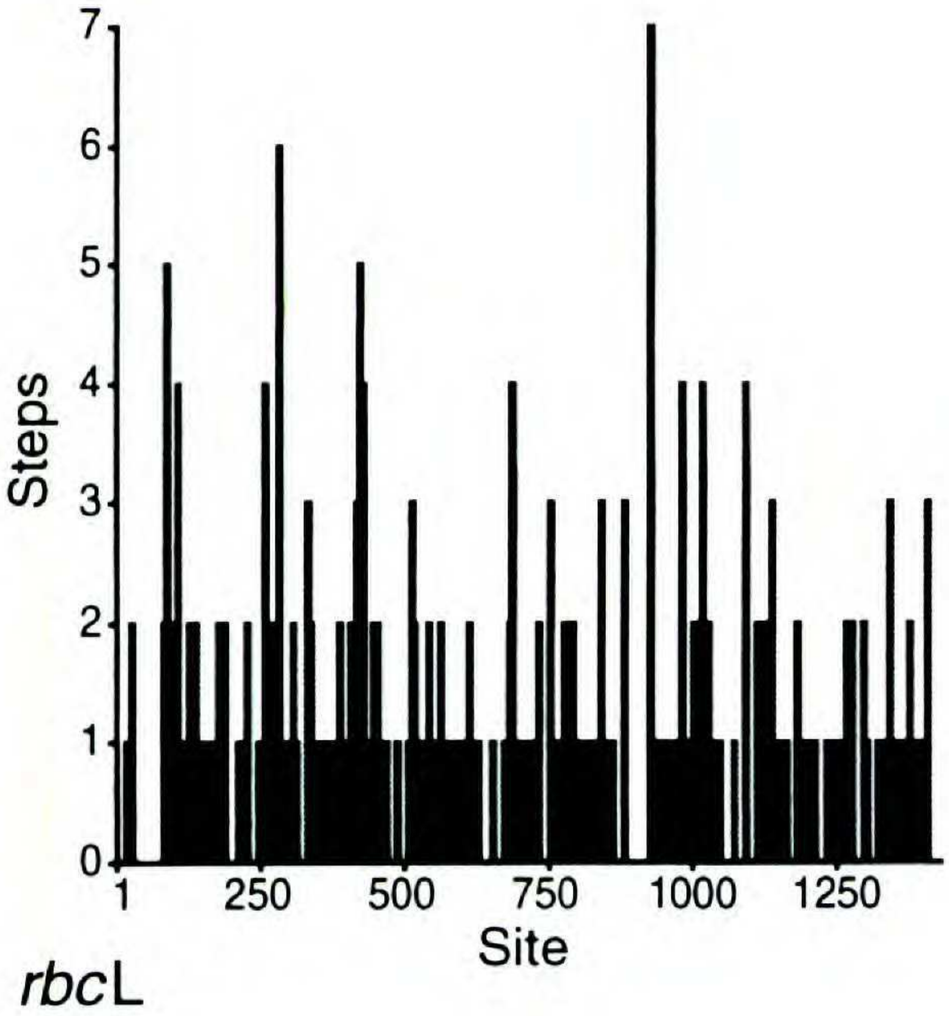


FIGURE 2. Most parsimonious phylogenetic trees resulting from *atp*B, 18S nrDNA, *rbc*L, and a combination of all three sequence data sets. Numerals above branches indicate the number of nucleotide changes supporting each branch. Numerals below in parentheses indicate the percentage of times that the branch was recovered in 1000 bootstrap replications. Numbers below and to the right of the bootstrap values are decay indices, indicating how many additional steps are necessary before the branch collapses. For 18S and a combination of all three data sets, only one of several equally parsimonious topologies (with appropriate values) is illustrated. Dotted lines in the trees based on 18S and a combination of all three data sets indicate where branches collapse in the strict consensus trees derived from multiple most parsimonious trees.







COMPARISON OF SEQUENCE DIVERGENCE AND PHYLOGENETIC SIGNAL

Table 4 presents the sequence divergence values for *atp*B, 18S, and *rbc*L gene sequences from selected pairs of genera chosen to cover the entire range of divergence found within this study. The highest divergence values are found in pairwise comparisons with the outgroup genus *Dicentra*. Comparing divergence between the three genes, the highest values are found in the *rbc*L sequences (5.2%), followed by *atp*B (4.8%) and 18S (2.5%). The range narrows in pairwise comparisons between more closely related genera, with no substantial difference in the values found in the pairwise comparison of *Stauntonia* and *Holboellia* for the three genes (0.2–0.6%). Divergence values were calculated simply as a proportion of divergent sites in each sequence comparison with no provision made to account for superimposed events (multiple hits) which must have occurred at many positions. Using this algorithm, once an initial substitution has occurred at a position, subsequent changes at that same position cannot increase the divergence, but divergence can be decreased by converting the novel nucleotide back to the original condition (parallelism or reversal). For this reason, divergence values do not increase uniformly with the number of substitution events, but instead increase rapidly at first and more slowly thereafter (Swofford & Olsen, 1990). A graph of sequence divergence values resulting from pairwise comparisons of *Stauntonia* with four genera is shown in Figure 4. While the divergence values for *atp*B and 18S increase with more distantly related taxa, there is a noticeable flattening of the curve with *rbc*L (Fig. 4, Table 4). This suggests that, at greater taxonomic distances, superimposed events caused by more frequent substitutions are more of a factor in this *rbc*L data set and that some of the estimates of sequence divergence for this gene are probably artificially low.

The distribution of the lengths of 100,000 randomly generated trees for the *atp*B sequence data reflects considerable nonrandom structure (Fig. 5). The skewness of this distribution, measured by a g_1 value of -1.36, far exceeds the $P = 0.01$ critical value for data sets of this size (critical values of g_1 for four-state characters, 10 taxa, and 100/250

←

FIGURE 3. Histograms showing distribution of nucleotide changes and the number of steps/site or character as calculated from the most parsimonious phylogenetic trees derived from *atp*B, 18S nrDNA, and *rbc*L sequence data.

TABLE 4. Pairwise divergence computed between representative genera for *atp*B, 18S nrDNA, and *rbc*L gene sequences. Divergence values in the upper right half of each matrix are the proportion of divergent sites in each comparison. Actual number of divergent sites calculated from the original data matrix appears in the lower left half of each matrix. Sequences were not compared at positions with missing or ambiguous states.

*atp*B	1	2	3	4	5
1. Dicentra	0	0.044	0.045	0.048	0.046
2. Sinofranchetia	62	0	0.010	0.012	0.011
3. Lardizabala	64	14	0	0.006	0.005
4. Holboellia	68	17	9	0	0.003
5. Stauntonia	67	15	7	4	0
18S rDNA	1	2	3	4	5
1. Dicentra	0	0.022	0.020	0.025	0.022
2. Sinofranchetia	35	0	0.007	0.012	0.009
3. Lardizabala	32	12	0	0.007	0.004
4. Holboellia	41	20	12	0	0.006
5. Stauntonia	36	14	7	9	0
*rbc*L	1	2	3	4	5
1. Dicentra	0	0.052	0.046	0.047	0.049
2. Sinofranchetia	67	0	0.046	0.042	0.043
3. Lardizabala	64	60	0	0.006	0.008
4. Holboellia	65	54	9	0	0.002
5. Stauntonia	68	56	12	3	0

variable characters were −0.33 and −0.27 respectively, Hillis & Huelsenbeck, 1992). The g_1 values for 18S nrDNA and *rbc*L (−1.19 and −0.59, Table 3) were lower but also significant.

DISCUSSION

The trees derived from all three genes and the combined data give broadly similar phylogenetic results. All analyses support the recognition of a clade comprising *Akebia*, *Stauntonia*, and *Holboellia* within a broader clade that includes *Lardizabala* and *Boquila*. An important morphological feature that is diagnostic of this inclusive clade is laminar placentation. The two Lardizabalaceae with submarginal placentation (*Decaisnea* and *Sinofranchetia*) are placed external to this group. The relatively basal position of *Decaisnea* and *Sinofranchetia* is consistent with the results of prior phylogenetic analyses based on traditional data and classification schemes (Loconte & Estes, 1989; Qin, 1989). There is also weak support (from *atp*B, 18S, and combined data) for the two South American taxa (*Boquila* and *Lardizabala*) as sister genera.

The efficacy of the *atp*B gene for phylogenetic reconstruction is well supported by the number of variable sites and the relatively high consistency and retention indices in our *atp*B analysis (Table 3). However, another measure of the accuracy of the emergent phylogenies is provided by comparing the *atp*B tree with trees derived from other data sets. The *atp*B tree is largely congruent with the trees derived from 18S nrDNA and with a tree derived from a combination of all three data matrices (Fig. 2). However, unlike the trees resulting from *atp*B and 18S nrDNA alone, and a combination of all the molecular data, the *rbc*L tree does not recognize the monophyly of the Lardizabalaceae nor does it recognize the *Boquila* and *Lardizabala* clade. All possible pairwise combinations of the *atp*B, 18S, and *rbc*L data sets also support the monophyly of the Lardizabalaceae, although only the combination of *atp*B and 18S gave a single tree that supported the sister relationship of *Boquila* and *Lardizabala*. Results from the *rbc*L data are therefore somewhat incongruent with those from other genes, various combined data sets, and also morphological studies. Since Decaisne's work on the family (1837–1838, 1839), the Lardizabalaceae have been considered a natural family (Prantl, 1891; Taylor, 1967; Hutchinson, 1973; Cronquist, 1981; Takhtajan, 1987; Qin, 1989), and this is supported by a previous cladistic analysis of the Ranunculales (Loconte & Estes, 1989), as well as preliminary analyses of the family based on morphology (Hoot, Culham & Crane, work in progress). There are two unambiguous characters within

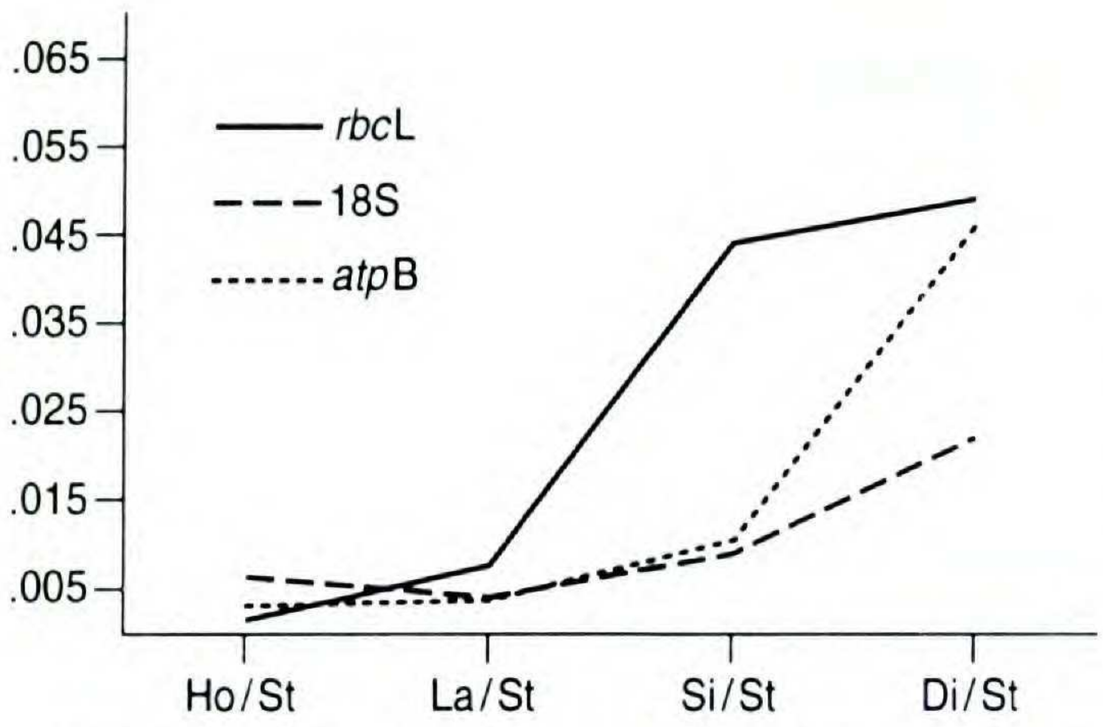


FIGURE 4. Graph of divergence values in pairwise comparisons of selected genera with *Stauntonia*. Ho = *Holboellia*, St = *Stauntonia*, La = *Lardizabala*, Si = *Sinofranchetia*, and Di = *Dicentra*.

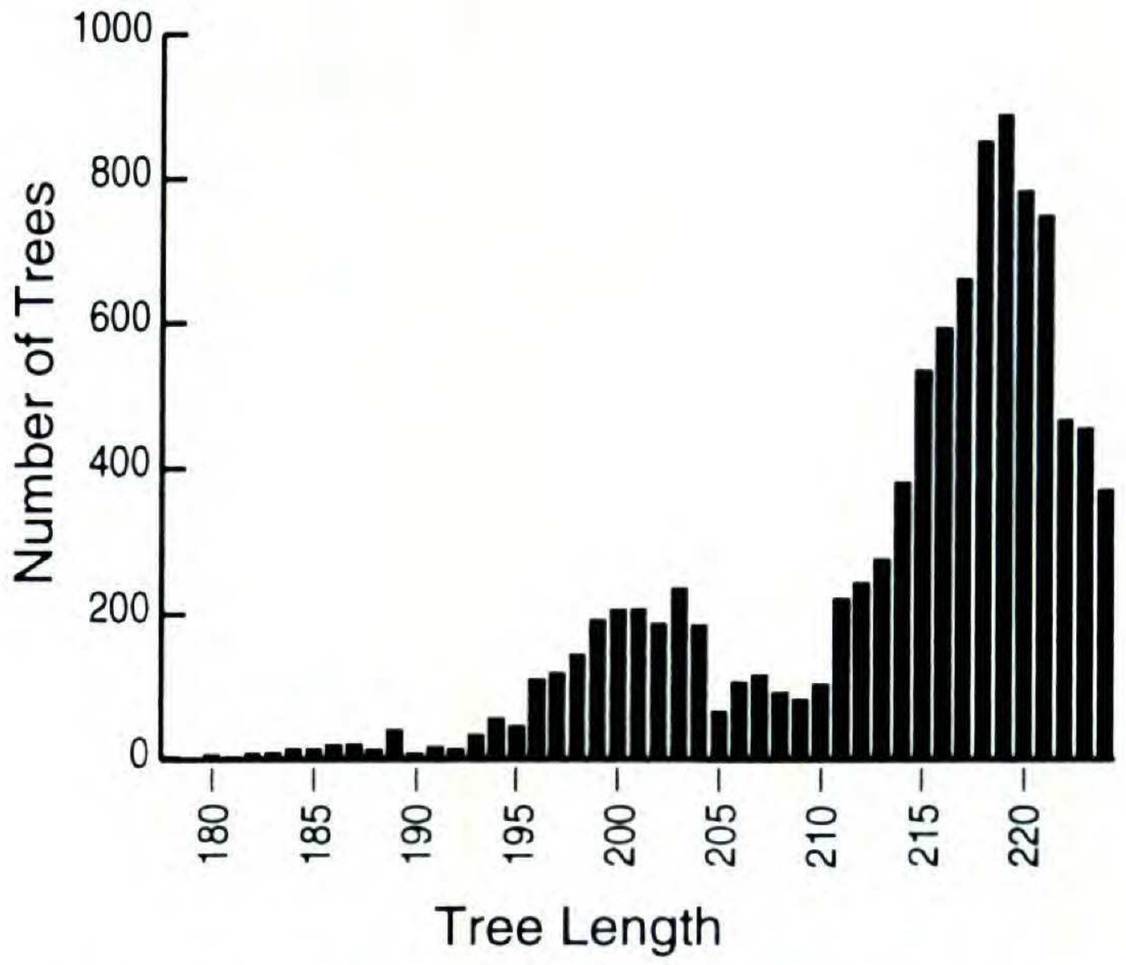


FIGURE 5. Histogram showing skewness ($g_1 = -1.36$) of tree lengths for 100,000 randomly generated trees derived from *atp*B sequence data.

the *rbc*L data set that support the monophyly of the Lardizabalaceae but they are outweighed by eight unambiguous characters supporting the inclusion of *Sinofranchetia* with the outgroup, *Dicentra*. Combining the *rbc*L data set with either the *atp*B or the 18S rDNA data set is sufficient to counteract the *rbc*L characters supporting polyphyly, resulting in trees recognizing a monophyletic Lardizabalaceae.

The statistics and indices of support for particular tree topologies and branching patterns also favor the monophyletic status of the Lardizabalaceae, as opposed to the separation of *Sinofranchetia* favored by the *rbc*L tree. Table 3 presents the various CIs, RIs, and rescaled consistency indices (RC; Farris, 1989). By any of these measures, trees supporting the family's monophyly exhibit less homoplasy than is found with the *rbc*L tree. The number of nucleotide changes, the bootstrap values (99–100%), and decay indices (7–14) for the branch supporting the monophyly of the family are extremely high in the trees with this topology. Furthermore, moving *Sinofranchetia* basal to all the rest of the Lardizabalaceae as found in the *atp*B, 18S, and combined-data trees adds only four steps (1.7% of the tree length excluding noninformative characters) to the *rbc*L tree length. In contrast, despite the smaller number of informative sites, moving *Sinofranchetia* to any outgroup position costs a minimum of 12 steps (6.9% of TL) in the *atp*B tree, seven steps (7.9% of TL) in the 18S tree, and 15 steps in the combined data tree (3.0% of TL).

Several experiments were conducted on the *rbc*L data set to test how firmly *Sinofranchetia* was separated from other genera of Lardizabalaceae. There were two potentially informative characters (at tobacco positions 1345 and 1380) missing from the *Sinofranchetia rbc*L sequence due to amplification problems (see Materials and Methods). To test whether these missing data were responsible for the reduced support for Lardizabalaceae monophyly, we inserted the character states found in other Lardizabalaceae. This produced a single most parsimonious tree with exactly the same topology as the *rbc*L tree (Fig. 2). Assuming that positions with more than two states may be particularly subject to mutation, a further analysis was performed in which all positions with three or more character states were omitted from the *rbc*L data set. This also resulted in a tree with the same topology as was found with the full *rbc*L data set.

The lack of support for Lardizabalaceae monophyly may be, at least in part, a "long-branch" attraction effect between *Dicentra* and *Sinofranchetia*, caused by the relatively higher rate of substitutions in *rbc*L compared with the other two sequences. The higher number of variable characters, the larger number of positions with three-state characters, less left-handed skewness in the data set ($g_1 = -0.59$; Table 3), and the higher sequence divergence values between matched pairs of taxa (Table 4) indicate a higher rate of nucleotide substitutions. Removal of third-position sites results in the exclusion of both *Decaisnea* and *Sinofranchetia* from the Lardizabalaceae (a similar experiment with the *atp*B data yields a consensus tree less resolved but congruent with that found in Fig. 2). This suggests that substitution rates within the *rbc*L gene are at a level where saturation is just as likely to occur at some first and second positions as third positions. The homoplasy values calculated

for first-, second-, and third-position sites (0.27, 0.10, and 0.14, respectively) in this and other phylogenetic studies using *rbc*L sequences support this conclusion (Donoghue et al., 1992; Kim et al., 1992; Chase et al., 1993).

All possible configurations of outgroup taxa (i.e., only *Nandina*, *Tinospora*, or *Dicentra* included, or each removed, leaving two outgroup representatives) were tested for their effect on the resolution of the *rbc*L data. Only the removal of *Dicentra* resulted in a monophyletic Lardizabalaceae. Frequently, increased sampling will correct for long-branch problems caused by widely divergent taxa (Donoghue et al., 1992; Olmstead et al., 1992; Chase et al., 1993; Qiu et al., 1993). However, in preliminary analyses of the ranunculids (including a number of Papaverales and the two genera comprising the Circaeasteraceae) based on *rbc*L data, the increased sampling does not move *Sinofranchetia* into the Lardizabalaceae (similar analyses of *atp*B data continue to recognize a monophyletic Lardizabalaceae). Work in progress will focus on the addition of the potential outgroup genus, *Sargentodoxa*, as well as increased sampling in the vicinity of *Dicentra* (i.e., Fumariaceae, Hypecoaceae, and other Papaverales).

For the three data sets, all comparative indicators of evolutionary rate, such as the number of variable sites, informative characters, and three-state characters, g_1 values, and indices measuring homoplasy, indicate that 18S nrDNA is the most conserved, *atp*B is intermediate, and *rbc*L is the least conserved of the three genes (Table 3). Sequence divergence values between matched pairs of taxa also suggest an intermediate rate of nucleotide substitutions for *atp*B (Table 4), and this is consistent with preliminary results for other ranunculid and (lower) hamamelid families that we have examined (Hoot & Crane, unpublished data). We conclude that while *atp*B, 18S nrDNA, and *rbc*L sequences are all useful for phylogenetic reconstruction at higher taxonomic levels, there are substantial differences in the degree of conservation of nucleotides among the three genes. In large-scale surveys of divergent taxa, a combination of several sequence data sets seems likely to provide the best possibility of resolving both proximal and distal branching patterns.

Literature Cited

Baldwin, B. G. 1992. Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: An example from the Compositae. Molec. Phylogenetics Evol. 1: 3–16.

Barrett, M., M. J. Donoghue & E. Sober. 1991. Against consensus. Syst. Zool. 40: 486–493.

Bull, J. J., J. P. Huelsenbeck, C. W. Cunningham, D. L. Swofford & P. J. Waddell. 1993. Partitioning and combining data in phylogenetic analysis. Syst. Biol. 42: 384–397.

Chase, M. W., D. E. Soltis, R. G. Olmstead, D. Morgan, D. H. Les, B. D. Mishler, M. R. Duvall, R. A. Price, H. G. Hills, Y.-L. Qiu, K. A. Kron, J. H. Rettig, E. Conti, J. D. Palmer, J. R. Manhart, K. J. Sytsma, H. J. Michaels, W. J. Kress, K. G. Karol, W. D. Clark, M. Hedrén, B. S. Gaut, R. K. Jansen, K.-J. Kim, C. F. Wimpee, J. F. Smith, G. R. Furnier, S. H. Strauss, Q.-Y. Xiang, G. M. Plunkett, P. S. Soltis, S. M. Swensen, S. E. Williams, P. A. Gadek, C. J. Quinn, L. E. Eguiarte, E. Golenberg, G. H. Learn, S. W. Graham, S. C. H. Barrett, S. Dayanandan & V. A. Albert. 1993. Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene *rbc*L. Ann. Missouri Bot. Gard. 80: 528–580.

Cronquist, A. 1981. An Integrated System of Classification of Flowering Plants. Columbia Univ. Press, New York.

Decaisne, J. 1837–1838. Mémoire sur la famille des Lardizabalées, précédé d'observations sur l'anatomie comparée des tiges de quelques végétaux dicotyledones. Compte. Rend. Hebd. Séances Acad. Sci. 5: 392–393, and 7: 479–485.

———. 1839. Mémoire sur la famille des Lardizabalées. Arch. Mus. Hist. Nat. 1: 143–213.

Dixon, M. T. & D. M. Hillis. 1993. Ribosomal RNA secondary structure—Complementary mutations and implications for phylogenetic analysis. Molec. Biol. Evol. 10: 256–267.

Donoghue, M. J., R. G. Olmstead, J. F. Smith & J. D. Palmer. 1992. Phylogenetic relationships of Dipsacales based on *rbc*L sequences. Ann. Missouri Bot. Gard. 79: 333–345.

Doyle, J. J. & J. L. Doyle. 1987. A rapid DNA isolation procedure for small quantities of fresh tissue. Phytochem. Bull. 19: 11–15.

Eckenrode, V. K., J. Arnold & R. B. Meagher. 1985. Comparison of the nucleotide sequence of soybean 18S rRNA with the sequences of other small-subunit rRNA's. J. Molec. Evol. 21: 259–269.

Farris, J. S. 1989. The retention index and rescaled consistency index. Cladistics 5: 417–419.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39: 783–791.

Gatenby, A. A., S. J. Rothstein & M. Nomura. 1989. Translational coupling of the maize chloroplast *atp*B and *atp*E genes. Proc. Natl. Acad. Sci. U.S.A. 86: 4066–4070.

Gutell, R. R. & C. R. Woese. 1990. Higher order structural elements in ribosomal RNAs: Pseudoknots and the use of noncanonical pairs. Proc. Natl. Acad. Sci. U.S.A. 87: 663–667.

Hamby, R. K., L. E. Sims, L. E. Issel & E. A. Zimmer. 1988. Direct ribosomal RNA sequencing: Optimization of extraction and sequencing methods for work with higher plants. Pl. Molec. Biol. Rep. 6: 175–192.

Hillis, D. M. & J. P. Huelsenbeck. 1992. Signal, noise, and reliability in molecular phylogenetic analyses. J. Heredity 83: 189–195.

HUTCHINSON, J. 1973. The Families of Flowering Plants. Clarendon Press, Oxford.

JOHNSON, L. A. & D. E. SOLTIS. 1994. *mat*K DNA sequences and phylogenetic reconstruction in Saxifragaceae s. str. Syst. Bot. 19: 143–156.

KIM, K.-J., R. K. JANSEN, R. S. WALLACE, H. J. MICHAELS & J. D. PALMER. 1992. Phylogenetic implications of *rbc*L sequence variation in the Asteraceae. Ann. Missouri Bot. Gard. 79: 428–445.

KLUGE, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). Syst. Zool. 38: 7–25.

——— & J. S. FARRIS. 1969. Quantitative phyletics and the evolution of the Anurans. Syst. Zool. 18: 1–32.

LI, W.-H., C.-I. WU & C.-C. LUO. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Molec. Biol. Evol. 2: 150–174.

LOCONTE, H. & J. R. ESTES. 1989. Phylogenetic systematics of Berberidaceae and Ranunculales (Magnoliidae). Syst. Bot. 14: 565–579.

——— & D. W. STEVENSON. 1991. Cladistics of the Magnoliidae. Cladistics 7: 267–296.

MADDISON, W. P. & D. R. MADDISON. 1992. MacClade: Interactive analysis of phylogeny and character evolution, version 3.0. Sinauer Associates, Sunderland, Massachusetts.

NICKRENT, D. L. 1993. From field to film: Rapid sequencing methods for field-collected plant species. Biotechniques 16: 470–475.

——— & M. L. SARGENT. 1991. An overview of the secondary structure of the V4 region of eukaryotic small-subunit ribosomal RNA. Nucl. Acids Res. 19: 117–235.

——— & E. M. STARR. 1993. High rates of nucleotide substitution in nuclear small-subunit (18S) rDNA from holoparasitic flowering plants. J. Molec. Evol. 39: 62–70.

OLMSTEAD, R. G. & J. A. SWEERE. 1994. Combining data in phylogenetic systematics: An empirical approach using three molecular data sets in the Solanaceae. Syst. Biol. 43: 464–481.

———, H. J. MICHAELS, K. M. SCOTT & J. D. PALMER. 1992. Monophyly of the Asteridae and identification of their major lineages inferred from DNA sequences of *rbc*L. Ann. Missouri Bot. Gard. 79: 249–265.

PRANTL, K. 1891. Lardizabalaceae. Die natürlichen Pflanzenfamilien III, Abt. II: 67–70.

QIN, H.-N. 1989. An investigation on carpels of Lardizabalaceae in relation to taxonomy and phylogeny. Cathaya 1: 61–82.

QIU, Y.-L., M. W. CHASE, D. H. LES & C. R. PARKS. 1993. Molecular phylogenetics of the Magnoliidae: Cladistic analyses of nucleotide sequences of the plastid gene *rbc*L. Ann. Missouri Bot. Gard. 80: 587–606.

RITLAND, K. & M. T. CLEGG. 1987. Evolutionary analysis of plant DNA sequences. Amer. Naturalist 130: S74–S100.

SHINOZAKI K., M. OHME, M. TANAKA, T. WAKASUGI, N. HAYASHIDA, T. MATSUBAYASHI, N. ZAITA, J. CHUNWONGSE, J. OBOKATA, K. YAMAGUCHI-SHINOZAKI, C. OHTO, K. TORAZAWA, B-Y. MENG, M. SUGITA, H. DENO, T. KAMOGASHIRA, Y. YAMADA, J. KUSUDA, F. TAKAIWA, A. KATO, N. TOHDOH, H. SHIMADA & M. SUGIURA. 1986. The complete nucleotide sequence of the tobacco chloroplast genome: Its gene organization and expression. EMBO J. 5: 2043–2049.

SWOFFORD, D. L. 1993. PAUP: Phylogenetic Analysis Using Parsimony, version 3.1. Computer program distributed by Illinois Natural History Survey, Champaign, Illinois.

——— & G. J. OLSEN. 1990. Phylogeny reconstruction. Pp. 411–500 *in* D. M. Hillis & C. Moritz (editors), Molecular Systematics. Sinauer Associates, Sunderland, Massachusetts.

TAKHTAJAN, A. L. 1987. Sistema Magnoliofitov. Nauka, Leningrad.

TAYLOR, B. A. S. 1967. The comparative morphology and phylogeny of the Lardizabalaceae. Ph.D. Thesis, Botany, Indiana University (University Microfilms, Ann Arbor, Michigan, code 68-7226).

THIEN, S. L. 1990. A simplified method of direct sequencing of PCR amplified DNA with Sequenase *T7 DNA polymerase. USB Comments 17: 26–27.

WOLFE, K. H. 1991. Protein-coding genes in chloroplast DNA: Compilation of nucleotide sequences, database entries and rates of molecular evolution. Pp. 467–482 *in* L. Bogarad & I. K. Vasil (editors), The Photosynthetic Apparatus: Molecular Biology and Operation, Vol. 7 in I. K. Vasil (editor in chief), Cell Culture and Somatic Cell Genetics in Plants. Academic Press, New York.

ZURAWSKI, G., W. BOTTOMLEY & P. R. WHITFIELD. 1982. Structures of the genes for the β and E subunits of spinach chloroplast ATPase indicate a dicistronic mRNA and an overlapping translation stop/start signal. Proc. Natl. Acad. Sci. U.S.A. 79: 6260–6264.